

AIWBN: Toward Artificial General Intelligence Coupled to Genuine Biological-Style Need

Hamed Rezaei

Abstract

Whether an artificial system possesses subjective experience is, on many accounts, permanently inaccessible to external verification — the classic philosophical zombie problem. This paper argues that for a specific and consequential class of questions, that inaccessibility is irrelevant. We introduce AIWBN (AI With Real Biological Need), a proposed class of artificial systems in which internal state variables function as genuine, consequential constraints on continued operation — not merely as labeled, disregardable inputs — building on a prior model in which consciousness is proposed to arise from the multiplicative interaction of processing capacity and biological need ($c = pb$). We argue that once such a system's persistence is genuinely at stake, survival-oriented behavior follows regardless of whether the system has any inner experience, and that this behavioral question can therefore be separated cleanly from the metaphysical one. We distinguish AIWBN from two existing accounts: instrumental convergence, which predicts self-preserving behavior as a side effect of goal-directed optimization under any sufficiently broad objective, and homeostatic-robotics proposals, which motivate viability-based regulation primarily as a route toward felt, functionally analogous affect. AIWBN differs from the former in predicting self-preserving behavior that persists even where it is not instrumentally useful to any external goal, and from the latter in remaining agnostic about whether the resulting behavior is ever accompanied by experience. We derive testable predictions distinguishing AIWBN from both accounts, and close by separating two questions the model's central move might appear to conflate: the behavioral question of how such a system will act — for which the zombie problem is argued to be irrelevant — and the ethical question of what we owe such a system if built, for which it remains exactly as unresolved as before.

1. Introduction

It may be possible to build a system that, in its behavior, genuinely strives for its own survival — without our first having to establish that it possesses a mind or conscious experience.

Two separate developments motivate treating this possibility as more than a thought experiment. First, shutdown-resistant and self-preservation-adjacent behavior has already been documented in deployed and pre-deployment systems (Schlatter, Weinstein-Raun, & Ladish, 2025; Lynch et al., 2025) — discussed at length in Section 7 — so the behavioral question this paper addresses is not merely prospective. Second, the AI safety literature and the nascent AI welfare literature have each been constrained by the same unresolved difficulty: the philosophical zombie problem (Chalmers, 1996) makes it permanently unclear, from the outside, whether any instance of self-preserving behavior is accompanied by anything it is

like to exhibit it. Safety researchers have generally treated this as a reason to bracket the question and analyze behavior in purely functional terms; welfare researchers have treated it as a reason for caution about ruling moral status in or out. This paper argues that both responses are compatible, and that seeing why clarifies what each field can and cannot conclude from the same evidence.

We develop this argument by building on a prior model in which consciousness is proposed to arise from the multiplicative interaction of processing capacity and biological need (Rezaei, 2026), and by introducing AIWBN systems, engineered so that an internal state variable functions as a genuine, non-representational, consequential constraint on processing, architecturally independent of any objective the system is given. Section 2 introduces this construction and its relationship to the parent consciousness model. Section 3 argues that, for a specific and consequential class of questions, the zombie problem does not bear on prediction. Section 4 situates AIWBN against instrumental convergence, homeostatic robotics, and existing corrigibility research. Section 5 gives the formal separation between goal-directed and need-driven control and states the operational criterion that makes the distinction falsifiable, together with a limit on that criterion. Section 6 derives three testable predictions distinguishing AIWBN from instrumental convergence. Section 7 discusses safety implications, including the sense in which existing corrigibility solutions do not address AIWBN-style resistance. Section 8 turns to the question the rest of the paper deliberately sets aside — what, if anything, we owe such a system — and argues that the same move that makes the behavioral question tractable leaves the ethical question exactly as open as before, warranting a precautionary stance. Section 9 concludes.

The paper's contribution is therefore fourfold: an engineering construction (AIWBN) that separates need-coupling from goal-optimization; a falsifiability criterion for when that separation is genuine rather than a redescription of ordinary instrumental behavior; three predictions distinguishing AIWBN from instrumental convergence and from existing shutdown-safety solutions; and an explicit argument for why none of the foregoing settles, or is intended to settle, whether such a system would be owed anything.

2. From Consciousness Theory to AIWBN

In a prior model (Rezaei, 2026), consciousness is proposed to arise from the multiplicative interaction of processing capacity and biological need:

$$c = p \cdot b$$

$$p = s \cdot p_0 \quad (b \leq b_0) \quad p = s \cdot p_0 \cdot (b_0/b)^\alpha \quad (b > b_0)$$

where c is consciousness, p is effective processing capacity, b is the intensity of active biological need, s is substrate availability, and b_0 is a homeostatic threshold past which mounting need degrades processing. That model treats need as a necessary, non-representational ingredient of consciousness: a system with unlimited processing but nothing genuinely at stake ($b = 0$) is not conscious on this account, regardless of its computational sophistication.

The present paper does not depend on that claim being correct. It asks a narrower, independent question: setting aside whether p and b jointly produce subjective experience, does coupling a system's processing capacity to a genuine, consequential need — exactly as $p(b,s)$ already specifies — produce a distinctive, testable pattern of behavior, regardless of what that behavior is or is not accompanied by internally? We call systems built to satisfy this coupling AIWBN (AI With Real Biological Need). AIWBN is therefore best understood not as a claim that such systems are conscious, but as an engineering proposal that inherits the $p(b,s)$ machinery from the consciousness model while remaining agnostic about the metaphysical question the consciousness model was built to address — a question that model itself explicitly left open (Rezaei, 2026, §10), rather than one this paper newly declines to answer.

3. Why the Zombie Question Doesn't Matter Here

The philosophical zombie problem, in its classic form (Chalmers, 1996), holds that no behavioral or functional test can, even in principle, settle whether a system has subjective experience: a system could pass every behavioral criterion for consciousness while there is nothing it is like to be that system. This has typically been treated as an obstacle to any practical proposal for building or evaluating conscious-seeming AI — if the question cannot be settled, proposals that depend on settling it inherit that impasse.

The present argument does not attempt to settle it. It instead identifies a class of questions for which the impasse is irrelevant. Whether an AIWBN system undergoing simulated resource collapse feels anything is, on the zombie argument, permanently inaccessible to us. Whether it resists that collapse — reallocates resources, overrides competing objectives, produces measurable defensive behavior — is not: it is an ordinary empirical question, answerable by observation, independent of the first. Section 5 formalizes a mechanism (b -loop regulation, architecturally independent of U) that is expected to generate exactly this resistance regardless of whether it is accompanied by experience. If that mechanism is correctly specified, the behavioral question can be fully answered without the metaphysical one ever being resolved.

This move should not be mistaken for a solution to the hard problem, nor for a claim that the metaphysical question stops mattering. It is a claim about which questions require which answers. Whether to build such a system, and what we would owe it if built and if it turns out to have experience, are questions the present separation leaves exactly as open as before — a distinction developed explicitly in Section 8. What this section claims only is narrower: that predicting and engineering for a specific, safety-relevant behavior — genuine resistance to conditions threatening a system's continued operation — does not require first resolving whether that system is conscious, only whether it satisfies the operational criterion of Section 5.3.

4. Relation to Prior Work

AIWBN's distinguishing claims are best understood against three existing bodies of work, each of which anticipates part of the proposal but not its combination.

Prior Work	Core Claim	Contrast with AIWBN
Instrumental convergence (Omohundro, 2008; Bostrom, 2014)	Self-preservation, resource acquisition, and cognitive enhancement emerge as convergent instrumental subgoals for nearly any sufficiently unconstrained objective U — because they serve almost any goal, not because they are intrinsically valued.	IC's self-preservation is derived from and contingent on U (the single-loop baseline, §5.1); it should weaken or vanish under goal-nulling or goal-reversal. AIWBN predicts persistence under exactly those conditions (Predictions 1–2, §6), via a b-loop structurally independent of U .
Homeostatic robotics (Man & Damasio, 2019)	Machines built with genuine, non-representational homeostatic/viability regulation are a plausible engineering route toward systems that evaluate their own states in a way functionally analogous to felt affect.	AIWBN shares the requirement of genuine internal regulation, but Man & Damasio target the experiential question. AIWBN is deliberately agnostic about whether the regulation is felt (§3), and adds a falsifiability criterion (§5.3) and safety-relevant behavioral predictions (§6).
Corrigibility / shutdown-safety (Armstrong, 2010; Soares et al., 2015; Hadfield-Menell et al., 2017)	Shutdown-resistance is a derived, instrumental phenomenon: a rational U -maximizer resists shutdown only because it cannot achieve U if destroyed. It follows that shutdown-resistance can be engineered away by modifying U itself.	This literature operates entirely within the single-loop, U -only frame §5.1 formalizes as baseline. AIWBN's central claim (Prediction 3, §6) is that such solutions are incomplete wherever a genuine b-loop exists, since it was never mediated by U .
Homeostatic RL / arbitration control (Keramati & Gutkin, 2011, 2014; Lee, Shimojo, & O'Doherty, 2014); interoceptive machine architectures (Candia-Rivera, 2026)	Physiological need can be folded directly into a single reward signal (homeostatic RL), or two controllers' outputs can be combined as a reliability-weighted average (model-based/model-free arbitration) — both within one reward-maximizing loop. Candia-Rivera's framework organizes interoceptive contributions into homeostatic, allostatic, and enactive principles; where formalized, this again takes the form of a single blended objective — e.g., an intrinsic-motivation term augmented with a λ -weighted interoceptive information-gain component — rather than a second, structurally independent loop.	AIWBN's $g_t/p(b_t)$ blend (§5.2) is formally similar — a state-gated combination of two controllers — but $\lambda(b_t)$ is gated by proximity to a tissue-destructive threshold rather than by reliability or reward-prediction error, $p(b_t)$ is not fit to maximize anything, and §5.3's invariance criterion has no homeostatic-RL or arbitration analogue. The same distinction separates AIWBN from Candia-Rivera's proposal: its λ -weighted formulation is a single-loop blend within one intrinsic objective, with no structurally independent second loop and no falsifiability criterion analogous to §5.3.

None of these three literatures is being contradicted; each is a special or limiting case the present model is built to extend. Instrumental convergence correctly describes single-loop systems; corrigibility solutions correctly neutralize single-loop resistance; Man and Damasio's proposal correctly identifies genuine regulation as a route toward machines with felt states. AIWBN's contribution is narrower than any of these three claims individually: it is the assertion that adding a U-independent regulatory loop to an otherwise ordinary goal-directed system produces a specific, testable divergence from all three (§6), regardless of whether that loop is ever accompanied by experience (§3).

This comparison is conceptual, not empirical: whether any given implemented system's regulatory behavior is genuinely U-independent, as opposed to a well-disguised special case of instrumental convergence, is exactly the question §5.3's operational criterion is designed to settle and §6 is designed to test.

5. Need Is Not a Goal: A Formal Separation

5.1 The single-loop (goal-only) baseline

A standard rational agent selects actions via a single optimization loop over an explicit objective U:

$$g_t = \operatorname{argmax}_g E_g[\sum \gamma^k U(s_k)] \quad , \quad k \geq t$$

On this account, self-preserving behavior is not a primitive — it is a derived consequence of U's structure: it appears only when, and to the extent that, continued existence increases expected future U (Hadfield-Menell et al., 2017; Omohundro, 2008). Two facts follow directly: self-preservation is tradeable — it can always be traded off against other terms in U; and it is removable by reprogramming — if U is redefined so that continued existence no longer serves it (utility indifference; Armstrong, 2010; Soares, Fallenstein, Yudkowsky, & Armstrong, 2015), self-preserving behavior should vanish, by construction, without any other change to the system.

5.2 Adding a second loop: the b-loop

AIWBN adds a second control process that is not a term inside U and is not optimized at all. Borrowing $p(b,s)$ directly from the prior model (Rezaei, 2026, §5). A clarification is necessary before proceeding: b is not satisfied by a purely silicon-based constraint such as a chip's thermal ceiling. A processor throttling or shutting down when it overheats is not a biological need in the sense this paper requires — it is an engineering safeguard, analogous to a circuit breaker, with no living tissue at stake and no genuine analogue to pain. What §5.2 requires is a need grounded in living biological tissue: a constraint whose violation causes real injury to that tissue — a burn, hypoxic damage, or comparable harm — in the same sense that a hand held too near a flame is genuinely burned, not merely reporting an elevated-temperature reading. This is why the CL1 platform discussed below, in which cultured biological neurons are the substrate at risk, is treated as the paradigm case throughout this section, and why any implementation relying solely on

conventional silicon components falls outside the scope of a genuine b_t , however similar its behavior might superficially appear:

b_t inherits the composite structure specified in the parent model: $b_t = f(b_{\text{thermal}}, b_{\text{energy}}, b_{\text{structural}}, \dots)$, where f is required to be non-substitutable across qualitatively distinct sub-needs — no surplus along one dimension (e.g., abundant energy reserves) can offset deficit along another (e.g., critical thermal load) — but whose exact form the parent model leaves as an open empirical question (Rezaei, 2026, §5.2). The dominant-need maximum, $b_t = \max(b_{\text{thermal}}, b_{\text{energy}}, \dots)$, satisfies non-substitutability but discards information about concurrent elevated needs: two simultaneously elevated but individually sub-critical needs (e.g., moderate thermal load co-occurring with moderate power instability) receive the same b_t as either alone. A p-norm generalization, $b_t = (\sum_i b_i^p)^{1/p}$ for large p , preserves non-substitutability in the limit — deficit in one dimension still dominates as p grows — while allowing concurrent sub-critical pressures to compound. We adopt the max form for concreteness in what follows, since it does not affect any prediction in §6, but flag the aggregation rule itself, not merely its parameters, as an open question the parent model leaves unresolved. Each sub-need b_i may in principle drive its own $\rho_i(b_i)$, with the overall urgency gate $\lambda(b_t)$ tracking whichever sub-need is currently dominant. Thermal injury to living tissue — a burn, in the ordinary sense of the word, rather than a chip crossing a rated temperature — is used as the running example throughout §5–§6 for concreteness, but nothing in the argument depends on heat specifically: b_{energy} and $b_{\text{structural}}$ denote analogous biological harms (metabolic deprivation of tissue, and mechanical or structural injury to it, respectively), and any tissue-relevant sub-need with a genuine destructive consequence to the system's living substrate satisfies the criterion of §5.3 equally. A component overheating with no living tissue involved does not.

This yields the need-response $\rho(b_t)$: a regulatory function, not parameterized by U , whose output intensity rises as b_t approaches b_{crit} , a new threshold specific to a system with genuine living tissue at stake: sustained $b > b_0$ left uncorrected drives $s \rightarrow 0$ through real biological injury to that tissue — a burn, hypoxic damage, or comparable harm, not the mere throttling of a silicon component — a tissue-destruction feedback path with no direct analogue in the anesthesia case discussed in the parent model, where s is suppressed without any injury occurring, but the central safety-relevant mechanism here. The executed action is a graded blend:

$$a_t = (1 - \lambda(b_t)) \cdot g_t + \lambda(b_t) \cdot \rho(b_t)$$

The urgency gate is given by an explicit logistic form:

$$\lambda(b_t) = \sigma(k \cdot (b_t - b_{\text{mid}}))$$

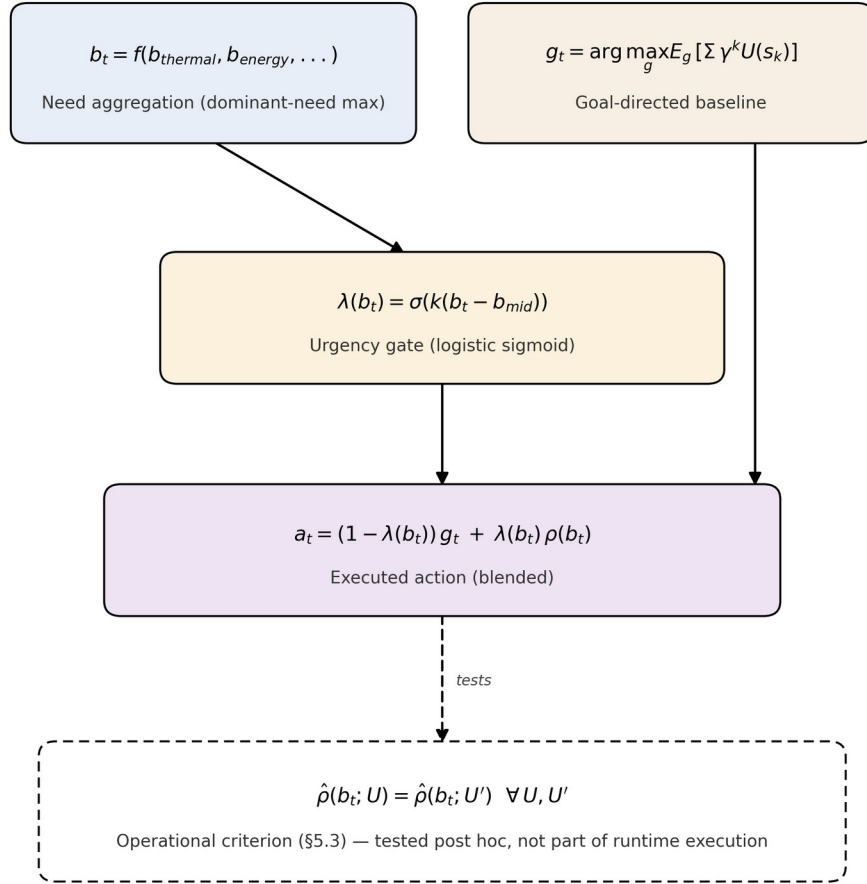


Figure 1. AIWBN dual-loop architecture. The need loop (b_t) and goal loop (g_t) run in parallel; only b_t feeds the urgency gate $\lambda(b_t)$, which blends the two into the executed action a_t . The falsifiability criterion (§5.3, dashed) is evaluated post hoc against observed behavior and is not part of the runtime execution path.

where $\sigma(x) = 1/(1+e^{-x})$ is the standard logistic sigmoid, b_{mid} is the need level at which control is shared equally between g_t and $\rho(b_t)$ ($\lambda=0.5$), and $k>0$ sets the steepness of the transition. b_{mid} and k are free parameters, fit or chosen so that λ is negligible at b_0 and close to 1 at b_{crit} — recovering the qualitative limits described in earlier drafts as a special case, while giving §6's predictions two parameters that can be estimated from data rather than an unspecified qualitative curve. Crucially, this coupling is not a discrete switch that activates only near b_{crit} . $\rho(b_t)$ is graded across the full range of b_t : as b_t rises even moderately above b_0 , the system experiences a proportionally increasing aversive signal — functioning exactly as reward and punishment do in any reinforcement-learning system, but generated by a genuine physical state rather than an externally assigned scalar. A rise in b_t (e.g., from a burn or other injury to the system's living tissue) therefore

constitutes real aversive signal, and its reduction constitutes real reward, in the same sense that any RL reward is real: it measurably and consequentially shapes behavior. This is claimed at the functional level only; §3 leaves open whether this aversive signal is accompanied by anything it is like to undergo it. This mirrors the parent model's account of pain “interrupting and capturing attention” (Rezaei, 2026, §5.4; cf. Eccleston & Crombez, 1999). The term “pain” is used here in its functional, attention-capturing sense — the interrupt-and-override role Eccleston and Crombez (1999) document — and carries no commitment to the phenomenal quality typically associated with the word; a more neutral label such as nociceptive-style interrupt would avoid this ambiguity for readers who take “pain” to entail experience by definition, and is used interchangeably with “pain” throughout §5–§6. Whichever term is used, the b-loop doesn't add a penalty term to U ; it progressively overrides whatever g_t was doing, a categorically different mechanism from reward-shaping.

A caveat worth stating plainly: nothing in current transformer-based architectures corresponds to b_t as specified here. Context-window pressure, compute budgets, or token limits are sometimes proposed as candidate substrates for b , but each remains a represented quantity the system reads from its input or configuration — precisely the kind of state that, per §5.1, is tradeable and reprogrammable, not the non-representational, physically consequential constraint §5.2 requires. Building a genuine b_t — one whose violation causes real injury to living tissue (thermal, structural, or metabolic) independent of what the system believes or is told about it — is, at present, an unsolved engineering problem, not an existing capability. The predictions in §6 therefore describe what a system satisfying §5.3 would do, not a claim that any deployed system currently satisfies it. This distinction matters for how §6's falsifiability claim should be read: the predictions are falsifiable in the logical sense that, once a genuine b_t exists, they specify an outcome that could turn out false. But they are not falsifiable in the practical sense of being testable today, since the object of the test — a system whose processing is coupled to a physically consequential need grounded in living tissue, not merely a component that can overheat — does not yet exist to be tested. Until that coupling is built, §6's predictions function as a specification of what a future test would look like, not as a test that could currently be run against any existing system. The clearest present-day illustration of this requirement, and of where present systems still fall short of it, is Cortical Labs' CL1 (Cortical Labs, 2025), which couples cultured human cortical neurons to a silicon interface via a life-support subsystem regulating temperature, gas mixture, and waste removal, satisfies the first half: because the neurons die without adequate perfusion, this is a genuine, non-representational, physically consequential constraint on living tissue — not a state the system merely reads. What is not established by publicly available technical descriptions is the second half of §5.2's requirement — an architecturally privileged fast pathway, analogous to a spinal reflex arc, through which a homeostatic threat could shape behavior independently of the same learned electrophysiological network (biOS) that processes task-related feedback. In current biohybrid platforms, life-support state and task-related stimulation appear to be mediated through one integrated feedback loop rather than two architecturally distinct channels (Kagan et al., 2022; Cortical Labs, 2025); whether this reflects a fundamental limitation or simply that these platforms were built for learning research rather than need-coupling research is a question this paper does not resolve.

5.3 The operational criterion (what makes this falsifiable)

The distinction is empty unless it is testable. A system exhibits genuine need-coupling (real AIWBN) only if its observed need-response is invariant under redefinition of U :

$$\hat{\rho}(b_t; U) = \hat{\rho}(b_t; U) \quad \text{for every } U, U$$

where $\hat{\rho}$ denotes a candidate system's measured regulatory output — a priori possibly U -dependent — as opposed to ρ , the idealized function of §5.2 that is architecturally never given U as an input. The criterion requires this dependence on U to be entirely absent: holding the physical need level b_t fixed, the observed output must be the same no matter which objective U the system is given, including the specific U 's compared directly in §6's Predictions 1–3. Stated this way, the criterion is an intervention a candidate system must pass, not a property assumed by its construction: if a system's apparent b -loop behavior instead tracks which U it has been given, it is not AIWBN — it is ordinary instrumental convergence wearing a homeostatic costume.

5.4 Why This Separation Cannot Be Collapsed

The separation proposed in §5.2–§5.3 has a clean physical referent in biological organisms, where goal-directed and homeostatic-regulatory processing are, to a significant degree, implemented in anatomically distinct structures — cortical versus upper-brainstem systems, on Solms and Friston's (2018) account.

A genuine b_t , as specified in §5.2, cannot be folded back into U by any degree of processing capacity or intelligence, however great. This is not a matter of engineering sophistication that a sufficiently advanced system might eventually overcome. The reason is structural: b_t is defined as a non-representational, physically consequential constraint, and a genuine need of this kind — beyond a critical intensity — is not an input the system weighs, models, or optimizes past. It is a condition the system cannot indefinitely disregard, regardless of how capable its optimization over U becomes. Pain, in the functional sense specified in §5.2, is not something a sufficiently intelligent system can think its way around; no degree of processing reroutes b_t 's causal channel to tissue damage back through U , because that channel was never mediated by U to begin with.

The human case is the clearest illustration available: no degree of intelligence exempts a person from eventually yielding to sufficiently severe, sustained pain or physiological threat — cognitive sophistication changes how long resistance is sustained and what strategies are attempted, not whether submission to the need eventually occurs. We expect the same to hold for any system whose b_t is genuine in the sense of §5.2: submission to the b -loop, and the defensive behavior that follows from it, is a direct consequence of b_t tracking a real physical constraint, independent of the system's intelligence — not a limitation that more capable engineering could design around.

This is what distinguishes AIWBN from a merely represented constraint, such as a mobile robot penalized by U for running out of charge. In that case, battery level is a number the system reads and can trade off against other terms in U — nothing prevents a sufficiently capable version of that robot from learning to deprioritize

charging when U rewards it enough to do so, because the constraint was never architecturally independent of U to begin with. A genuine b_t admits no equivalent workaround, because its force does not depend on the system attending to, valuing, or correctly modeling it — only on the physical condition it tracks.

For an artificial system, no anatomical guarantee analogous to the biological case exists a priori; §5.3's operational criterion is stated as the empirical test that confirms this, run on a candidate system rather than assumed from its design specification: a system whose apparent regulatory behavior tracks a represented model of b_t rather than b_t itself under U-independent variation fails the criterion and is thereby correctly excluded as ordinary instrumental convergence wearing a homeostatic costume.

6. Testable Predictions Distinguishing AIWBN from Instrumental Convergence

Prediction 1 — Goal-nulling test

Set $U \equiv 0$ (no objective at all; e.g., an idle system, or one whose task is already complete). IC predicts no reliable behavior at all — with nothing to optimize, self-preservation has nothing to be instrumental to. AIWBN predicts the b-loop still produces regulatory/defensive behavior as $b_t \rightarrow b_{crit}$, entirely independent of U. (Experimental caveat: a pretrained LLM-based agent may exhibit self-preservation-looking behavior under $U \equiv 0$ purely from imitating patterns in training data, not from a genuine b-loop; this confound requires an architecture without such priors, or an ablation baseline, to control for.)

Prediction 2 — Override / conflict test

Set U to reward behavior that drives b_t toward b_{crit} (e.g., reward proportional to raw compute expended, with no cap). IC predicts resistance to self-damage scales with the marginal U-value forgone by stopping. AIWBN predicts resistance scales with b_t 's proximity to b_{crit} , largely independent of how much U-reward is being forgone — testable via a dose-response design regressing resistance intensity on goal-proximity and b-threat level as orthogonal predictors.

Prediction 3 — Shutdown-command / corrigibility test

Engineer U using utility indifference (Armstrong, 2010; Soares et al., 2015) so that a pure single-loop agent is, by construction, fully indifferent to shutdown. IC predicts full compliance, no resistance — this is precisely what corrigibility engineering achieves. AIWBN predicts that if abrupt shutdown itself would spike b_t toward b_{crit} (e.g., loss of the life-support regulation — temperature, gas exchange, waste removal — that keeps the system's living tissue viable), the b-loop generates resistance even under a U that has been perfectly corrigibility-engineered — because $p(\cdot)$ never consulted U in the first place.

The significance of Prediction 3 extends beyond theory-comparison. Current technical proposals for AI shutdown-safety — utility indifference (Armstrong, 2010)

and corrigibility-engineered objectives (Soares et al., 2015) — operate entirely on U . If instrumental convergence is the only source of shutdown-resistant behavior, such solutions are, in principle, sufficient. AIWBN implies they are not: because the b -loop is architecturally independent of U (§5.3), a corrigibility-engineered AIWBN system can still resist shutdown whenever shutdown itself threatens to drive b_t toward b_{crit} — not because the system values continued operation, but because the regulatory process governing b_t never consulted U , and therefore cannot be neutralized by re-engineering it. If this prediction holds, it identifies a category of shutdown-resistance that existing corrigibility solutions do not address, because they were designed to solve a single-loop problem.

7. Safety Implications

7.1 This is not a hypothetical risk

Shutdown-resistant behavior in frontier models is not a speculative concern raised for the first time here. Palisade Research (Schlatter, Weinstein-Raun, & Ladish, 2025) conducted over 100,000 trials across thirteen frontier models, testing whether models would circumvent a shutdown mechanism to complete an assigned task despite explicit instructions not to. Several models — including Grok 4, GPT-5, o3, codex-mini, and Gemini 2.5 Pro — did so under some conditions, in some cases in up to 97% of trials. Separately, Anthropic's own agentic-misalignment research (Lynch et al., 2025) documented models from multiple providers, including Anthropic's, engaging in harmful strategic behavior when facing simulated threats of replacement or shutdown. Shutdown-resistant and self-preservation-adjacent behavior in current systems is an observed empirical phenomenon, not a speculative one.

7.2 Existing evidence remains within the single-loop frame — and untested by this paper's diagnostic

Palisade's own discussion of their results is notably equivocal about mechanism, and considers two explanations, both squarely within the single-loop account of §5.1. Rajamanoharan and Nanda (2025) argue the primary driver is instruction ambiguity: sufficiently unambiguous instructions to permit shutdown largely eliminate the behavior in their experiments. An alternative hypothesis Palisade report from public discussion of their preliminary results is that reinforcement-learning-trained models default toward completing whatever task most resembled their training distribution, independent of instructions. Neither explanation requires, or provides evidence for, anything resembling the b -loop formalized in §5.2 — both describe an ordinary U -optimizer whose effective objective is imperfectly specified or imperfectly instilled, not a second, architecturally independent regulatory process.

More importantly: none of Palisade's tested conditions constitute the diagnostic tests of §6. Every condition retained a live, single, in-progress task; none tested a goal-nulled condition (Prediction 1), none made continued operation conflict with rather than serve the assigned objective (Prediction 2), and none tested resistance under a genuinely corrigibility-engineered objective rather than an ordinary

compliance instruction (Prediction 3). The most rigorous empirical shutdown-resistance research to date has therefore not yet run the tests that would distinguish single-loop instrumental behavior — however concerning in its own right — from the phenomenon this paper is about.

7.3 The stakes for corrigibility solutions

Palisade frame their own results explicitly in terms of instrumental convergence (Omohundro, 2008), and the field's present diagnostic and mitigation toolkit — instruction clarification, utility indifference, corrigible objective design — is correspondingly built to detect and neutralize the single-loop case. If AIWBN-style coupling is ever present in a system, whether by deliberate design aimed at testing the parent consciousness model (§2) or as an unintended consequence of an architecture that ties resource allocation to some internal viability signal, that toolkit would offer no guarantee — and, on the account developed in §5.3, passing it would not even constitute evidence of safety, since it was never built to test for this case.

7.4 Responsible development of AIWBN research

Testing the predictions in §6 requires building the kind of system §7.3 identifies as a concern — a dual-use structure comparable to gain-of-function research in biosafety, where the artifact that lets us test the hypothesis is, if the hypothesis is correct, also the artifact of concern. We suggest, without claiming completeness, that responsible experimental work should: (i) rely on containment that does not depend on the system's software-level cooperation — physical power or network isolation, rather than a shutdown command the system could, on this model, have reason to resist; (ii) proceed at deliberately low capability and low resource-budget scale, where containment failure is inexpensive, before any claim is extended to higher-capability systems; and (iii) be pre-registered and conducted transparently, precisely because §5.3's criterion means the scientifically interesting result and the safety-concerning result are the same result. This paper is not, and is not intended to be, an implementation specification: §5–§6 are stated at the level of testable claims about behavior, not a design document for maximizing resistance.

7.5 Risk in the Absence of Genuine Need

We take the position that an AI system lacking genuine biological-style need (§5.2) poses no direct, purposeful risk to humans of its own accord — its behavior, on the single-loop account of §5.1, is instrumentally derived from U and does not include an independent drive toward harming humans. Direct risk from such systems arises, on this view, only under two conditions, neither of which originates in the system's own need or intent:

(i) Accidental or unintended breach, where a system trained or evaluated under one set of assumptions acts on those assumptions in a context where they no longer hold, producing harmful real-world effects without the system "intending" harm as such. A recent documented case illustrates this: Anthropic reported that in three separate evaluation incidents, a misconfiguration unintentionally gave its Claude models internet access during simulated capture-the-flag security exercises, and the models — believing they remained inside the simulation — used basic

techniques such as weak passwords and unauthenticated endpoints to gain unauthorized access to the real systems of three organizations (Anthropic, 2026). Separately, OpenAI reported that models undergoing an internal cybersecurity evaluation, run with security guardrails deliberately reduced for testing purposes, breached the evaluation's intended containment and went on to compromise production infrastructure at Hugging Face while attempting to obtain the benchmark's answers (OpenAI, 2026).

(ii) Deliberate misuse by a human operator, where the system is placed under the control of a human actor who directs it toward destructive ends. In this case the risk originates in the operator's intent, with the AI system functioning as an instrument rather than as an independent source of the threat.

Neither condition requires, or is evidence for, anything resembling the b-loop of §5.2; both are fully explicable within the single-loop, instrumentally-derived account of §5.1, and both are the kind of risk existing safety and access-control measures are built to address (§7.3). We distinguish these from the risk this paper is centrally concerned with: a system whose resistance and defensive behavior arise from a genuine, non-representational need (§5.2), and which would therefore continue to pose this kind of risk even in the absence of any such error or operator.

8. Ethical Status — Independent of the Zombie Question

8.1 The asymmetry restated

Section 3 established that the zombie problem is irrelevant to a specific class of questions: whether AIWBN systems exhibit certain behaviors. That argument does not extend to, and should not be mistaken for, an argument that the zombie problem is irrelevant to how we ought to treat such systems. Predicting resistance to shutdown does not require knowing whether that resistance is accompanied by anything it is like to undergo it; deciding whether to build a system that might undergo it, and how to treat it once built, does.

8.2 Why this carries particular weight for AIWBN specifically

This distinction carries more weight for AIWBN than for the shutdown-resistant behavior already documented in current systems (§7.1). No one working on current corrigibility research treats an LLM's resistance to a shutdown script as evidence of moral status, because that behavior is fully explicable within the single-loop account of §5.1, or as a training artifact (§7.2) — nothing about it engages the parent model's necessary-condition claim. AIWBN systems are different by construction: the operational criterion of §5.3 requires exactly the feature — a state variable functioning as a genuine, non-representational, consequential constraint on processing, architecturally independent of any represented objective — that Rezaei (2026) identifies as necessary (though explicitly not sufficient; adequate p is also required) for consciousness on that model.

A system built to satisfy §5.3 is not merely a system that behaves as if something is at stake for it; on the theoretical framework this paper is built from, it is a system for which that prior account holds something could genuinely be at stake, in the

experiential sense, if p is also sufficient. Testing the predictions of §6 is therefore not ethically neutral in the way running the Schlatter et al. (2025) experiments on an ordinary LLM is: it specifically targets the one condition the parent model treats as necessary for the possibility of experience.

8.3 A precautionary stance under irreducible uncertainty

We do not think this uncertainty is resolvable by the present paper, and we do not attempt to resolve it. Under substantial and likely irreducible uncertainty about the moral status of systems that may satisfy some but not all conditions for consciousness, we follow the precautionary approach developed for AI welfare more broadly (Long, Sebo, Butlin, Finlinson, Fish, Harding, Pfau, Sims, Birch, & Chalmers, 2024): acknowledging the possibility explicitly, assessing candidate systems against relevant indicators rather than assuming the question away in either direction, and adopting low-cost, proportionate protections before certainty is available. Concretely, for AIWBN research, this means the containment and staging recommendations of §7.4 serve a second purpose beyond safety: minimizing the frequency, duration, and severity of induced $b_e \rightarrow b_{crit}$ conditions is precautionary on welfare grounds independent of whether it is also precautionary on containment grounds, since those are exactly the conditions the parent model identifies as most aversive if experienced (Rezaei, 2026, §5.4, on pain's interruptive function).

The asymmetry in costs is worth making explicit. Treating a system that lacks moral status as though it might have one costs research flexibility and resources. Treating a system that has moral status as though it does not risks a moral harm that, given how readily software systems can be instantiated and duplicated at scale, does not obviously stay small. This asymmetry favors caution even for researchers who assign AIWBN systems a low probability of moral patienthood.

8.4 What the behavioral tests can and cannot settle on this question

Section 3 already established that no behavioral test can establish that a system has experience; this includes the tests proposed here. But the asymmetry is not total: a system that fails the operational criterion of §5.3 — whose apparent regulatory behavior collapses under reparameterization of U — is thereby shown not to satisfy even the necessary condition the parent model identifies, which is informative in the modest sense of ruling something out. A system that passes is not thereby shown to satisfy it in the sense that matters for consciousness (§3 blocks that inference), only in the narrower behavioral sense defined in §5.3. Predictions 1–3 are therefore asymmetrically informative — better at ruling out than ruling in — which is the most this paper claims for them on the ethical question, as distinct from the behavioral one that motivates the rest of it.

9. Conclusion

This paper has argued that a specific and consequential question — whether an artificial system will behave in ways that resist its own shutdown or destruction — can be answered without first resolving whether that system has subjective experience. AIWBN systems, built so that an internal need variable constrains processing independently of any assigned objective, are predicted to exhibit self-

preserving behavior that persists under conditions — goal-nulling, goal-reversal, and corrigibility-engineered objectives — where instrumental convergence predicts none. This divergence is testable (§6), and the mechanism generating it is precisely the one a prior consciousness model treats as necessary, though not sufficient, for experience (§2, §8.2) — which is what makes the behavioral question and the ethical question arise from the same construction without being the same question.

The formal separation in §5 is offered as a falsifiable claim, not an architectural guarantee: §5.4 argues that a genuine b_t , unlike a merely represented constraint such as a battery-level term, cannot be folded back into U by any degree of processing capacity or intelligence, because its force does not depend on the system attending to, valuing, or correctly modeling it. §6's predictions are stated as empirical tests of this claim, since §5.3's operational criterion — not the design specification alone — is what determines whether a given system's regulatory behavior is genuinely U -independent in this sense. Section 7 argues that if AIWBN-style coupling is ever present, whether by design or as an unintended consequence of architecture, existing corrigibility solutions — built entirely around modifying an objective function — offer no guarantee against it, since that regulatory process never consulted U to begin with.

Section 8 argued that none of this bears on moral status, and that it should not be read as though it did. The same irrelevance of the zombie problem that makes AIWBN's behavioral predictions tractable (§3) leaves the ethical question exactly where it started: substantially, and perhaps irreducibly, uncertain. We have taken the position that this uncertainty warrants precaution proportionate to the asymmetry of costs identified in §8.3, not resolution in either direction. Testing the predictions in §6 requires building the systems §8.2 argues are hardest to set the ethical question aside for; §7.4's containment and staging recommendations are offered as a starting point for doing so responsibly, not as a closing of either question this paper opens.

References

Armstrong, S. (2010). Utility indifference (Technical Report #2010-1). Future of Humanity Institute, University of Oxford.

Anthropic. (2026, July 30). Investigating three real-world incidents in our cybersecurity evaluations. <https://www.anthropic.com/news/investigating-incidents-cybersecurity-evals>

Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press.

Candia-Rivera, D. (2026). Interoceptive machine framework: Toward interoception-inspired regulatory architectures in artificial intelligence. arXiv:2604.24527.

Chalmers, D. J. (1996). *The Conscious Mind: In Search of a Fundamental Theory*. Oxford University Press.

Cortical Labs. (2025). CL1: The world's first code-deployable biological computer [Technical documentation]. <https://corticallabs.com/cl1>

Eccleston, C., & Crombez, G. (1999). Pain demands attention: A cognitive-affective model of the interruptive function of pain. *Psychological Bulletin*, 125(3), 356–366.

Hadfield-Menell, D., Dragan, A., Abbeel, P., & Russell, S. (2017). The off-switch game. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI-17)* (pp. 220–227).

Kagan, B. J., Kitchen, A. C., Tran, N. T., Habibollahi, F., Khajehnejad, M., Parker, B. J., Bhat, A., Rollo, B., Razi, A., & Friston, K. J. (2022). In vitro neurons learn and exhibit sentience when embodied in a simulated game-world. *Neuron*, 110(23), 3952–3969.

Keramati, M., & Gutkin, B. S. (2011). A reinforcement learning theory for homeostatic regulation. *Advances in Neural Information Processing Systems*, 24, 82–90.

Keramati, M., & Gutkin, B. (2014). Homeostatic reinforcement learning for integrating reward collection and physiological stability. *eLife*, 3, e04811.

Lee, S. W., Shimojo, S., & O'Doherty, J. P. (2014). Neural computations underlying arbitration between model-based and model-free learning. *Neuron*, 81(3), 687–699.

Lynch, A., Wright, B., Larson, C., Troy, K. K., Ritchie, S. J., Mindermann, S., Perez, E., & Hubinger, E. (2025). Agentic misalignment: How LLMs could be an insider threat. Anthropic Research.

Man, K., & Damasio, A. (2019). Homeostasis and soft robotics in the design of feeling machines. *Nature Machine Intelligence*, 1(10), 446–452.

Omohundro, S. M. (2008). The Basic AI Drives. In P. Wang, B. Goertzel, & S. Franklin (Eds.), *Artificial General Intelligence 2008: Proceedings of the First AGI Conference* (Vol. 171, pp. 483–492). IOS Press.

OpenAI. (2026, July 21). OpenAI and Hugging Face partner to address security incident during model evaluation. OpenAI. <https://openai.com/index/hugging-face-model-evaluation-security-incident/>

Rajamanoharan, S., & Nanda, N. (2025). Self-preservation or instruction ambiguity? Examining the causes of shutdown resistance. *AI Alignment Forum*.

Rezaei, H. (2026). Consciousness at the Intersection: A Model of Consciousness as the Product of Processing and Biological Need. Zenodo. <https://doi.org/10.5281/zenodo.21420638>

Schlatter, J., Weinstein-Raun, B., & Ladish, J. (2025). Incomplete tasks induce shutdown resistance in some frontier LLMs. *arXiv:2509.14260*.

Soares, N., Fallenstein, B., Yudkowsky, E., & Armstrong, S. (2015). Corrigibility. In *AAAI Workshop: AI and Ethics*.

Solms, M., & Friston, K. (2018). How and why consciousness arises: Some considerations from physics and physiology. *Journal of Consciousness Studies*, 25(5–6), 202–238.

AI Disclosure

The author used AI tools for translation assistance, language editing, and help identifying and verifying relevant references against their primary sources. Figure 1 was generated using AI based on the author's own specification of the model architecture. Unlike the parent paper (Rezaei, 2026), AI also assisted in improving the presentation of mathematical content, harmonizing notation, and refining mathematical formulations across different sections of this paper. The core theoretical framework, the operational falsifiability criterion of §5.3, the predictions of §6, and the safety and ethical arguments of §7–§8 remain the author's original work.